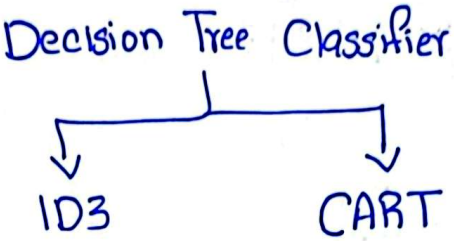
Decision Tree

—> Decision Tree Classifier



① Purity
- ↳ Entropy
- ↳ Gini Impurity

② What feature needs to select for splitting
- ↳ Information Gain

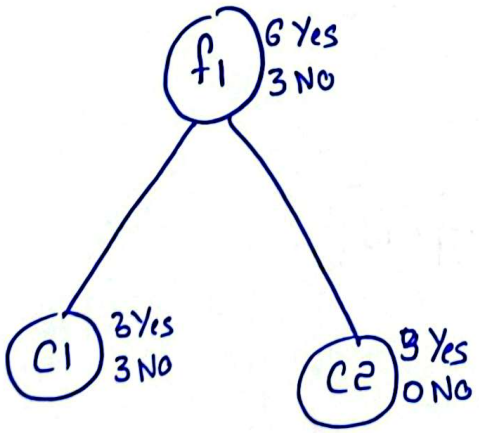
> Entropy and Gini Impurity

Entropy

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

Gini Impurity

$$Gini = 1 - \sum_{i=1}^{c} p_i^2$$



$$= 1 - [P(Y)^2 + P(N)^2]$$

$$= 1 - [(1/2)^2 + (1/2)^2]$$

$$= 0.5 \Rightarrow \text{Impure Split}$$

$$E(C_1) = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)$$

$$= 1$$

$$E(C_2) = -\frac{3}{3}\log_2(3/3) - 0\log_2 0$$

$$= -\log_2 1 = 0$$

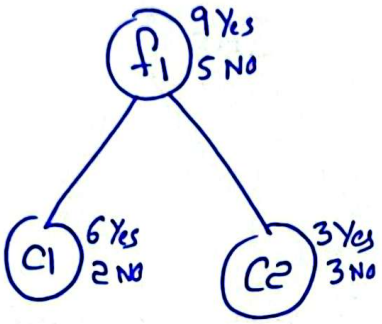
when dataset is small, use Entropy

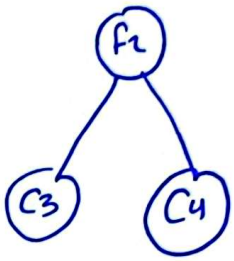
when dataset is large, use Gini Impurity

> Information Gain

$$Gain(S,f) = E(S) - \sum_{j=1}^{K} F(j, f_v)$$

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$





$$E(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$$

$$= 0.94$$

$$E(C_1) = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}$$

$$= 0.81$$

$$E(C_2) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}$$

$$= 1$$

$$Gain(S,f_1) = 0.94 - \left[\frac{8}{14}(0.81) + \frac{6}{14}(1)\right]$$

$$= 0.049$$

Suppose $Gain(S,f_2) = \underbrace{0.051}$

Choose for Splitting

> Example

Day	Outlook	Temperature	Humidity	Wind	Decision
1	Sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
3	overcast	hot	high	weak	Yes
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
6	rainfall	cool	normal	strong	No
7	overcast	cool	normal	~~weak~~ strong	Yes
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes
11	sunny	mild	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes
14	rainfall	mild	high	strong	No

Gini Outlook

Gini Index (sunny) $= 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.48$

Gini Index (outlook = rainfall) $= 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$

Gini Index (outlook = overcast) $= 1 - (4/4)^2 - (0/4)^2 = 0$

Gini (outlook) $= \frac{5}{14}(0.48) + \frac{4}{14}(0) + \frac{5}{14}(0.48) = 0.342$

Temperature

Gi (temperature = hot) $= 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0.5$

Gi (temperature = mild) $= 1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2 = 0.445$

Gi (temperature = cool) $= 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0.375$

Gini (temperature) $= \frac{4}{14}(0.5) + \left(\frac{6}{14}\right)(0.445) + \frac{4}{14}(0.375) = 0.439$

Humidity

Gi (humidity = high) $= 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = 0.489$

Gi (humidity = normal) $= 1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2 = 0.244$

Gini (humidity) $= \frac{7}{14}(0.489) + \frac{7}{14}(0.244) = 0.367$

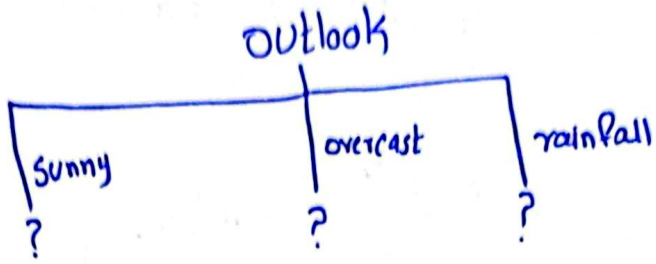
Wind

Gi (wind = strong) $= 1 - \left(\frac{6}{8}\right)^2 - \left(\frac{2}{8}\right)^2 = 0.375$

Gi (wind = weak) $= 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = 0.5$

Gini (wind) $= \frac{8}{14}(0.375) + \left(\frac{6}{14}\right)(0.5) = 0.428$

outbook has the lowest gini, it will be selected as root node



Outlook = Sunny

Gini(outlook = sunny and temperature = hot) $= 1-(0/2)^2-(2/2)^2 = 0$

Gini(outlook = sunny and temperature = cool) $= 1-(1/1)^2-(0/1)^2 = 0$

Gini(outlook = sunny and temperature = mild) $= 1-(1/2)^2-(1/2)^2 = 0.5$

Gini(outlook = sunny and temperature)

$= (2/5)(0) + (1/5)(0) + (2/5)(0.5) = 0.2$

Gini(outlook = sunny and humidity = high) $= 1-(0/3)^2-(3/3)^2 = 0$

Gini(outlook = sunny and humidity = normal) $= 1-(2/2)^2-(0/2)^2 = 0$

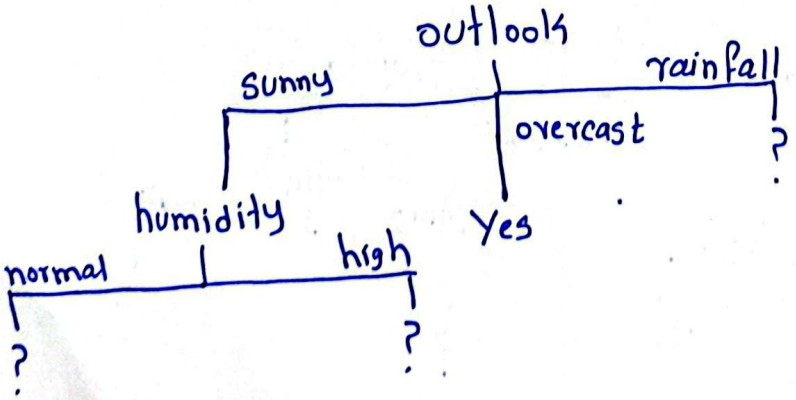
Gini(outlook = sunny and humidity)

$= 3/5(0) + 2/5(0) = 0$

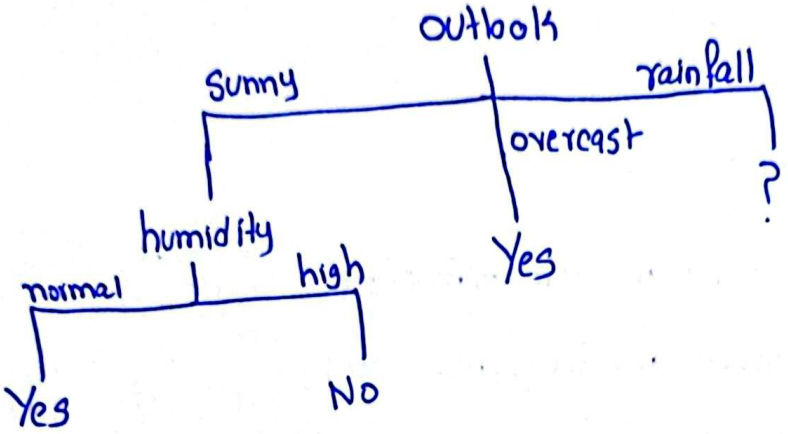
Gini(outlook = sunny and wind = weak) $= 1-(1/3)^2-(2/3)^2 = 0.44$

Gini(outlook = sunny and wind = strong) $= 1-(1/2)^2-(1/2)^2 = 0.5$

Gini(outlook = sunny and wind) $= 3/5(0.44) + (2/5)(0.5) = 0.466$



Lets focus on sub data for humidity



outlook = rainfall

Gini (outlook = rainfall and temperature = cool) = $1-(1/2)^2-(1/2)^2 = 0.5$

Gini (outlook = rainfall and temperature = mild) = $1-(2/3)^2-(1/3)^2 = 0.44$

Gini (outlook = rainfall and temperature) = $2/5(0.5) + 3/5(0.44) = 0.466$

Gini (outlook = rainfall and humidity = high) = $1-(1/2)^2-(1/2)^2 = 0.5$

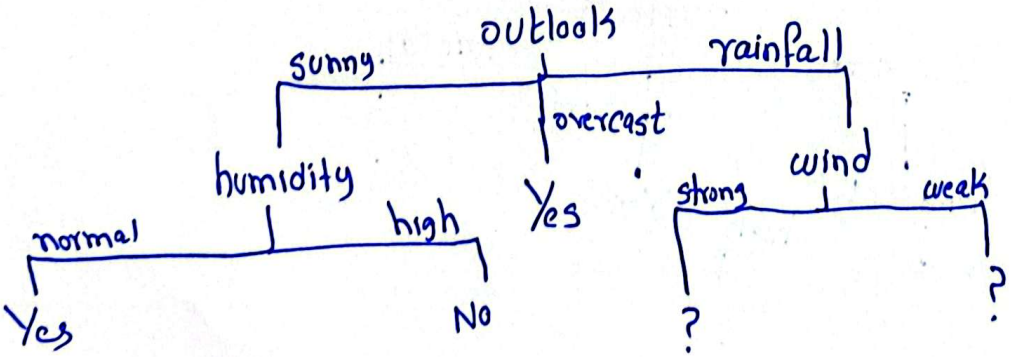
Gini (outlook = rainfall and humidity = normal) = $1-(2/3)^2-(1/3)^2 = 0.44$

Gini (outlook = rainfall and humidity) = $2/5(0.5) + 3/5(0.44) = 0.466$

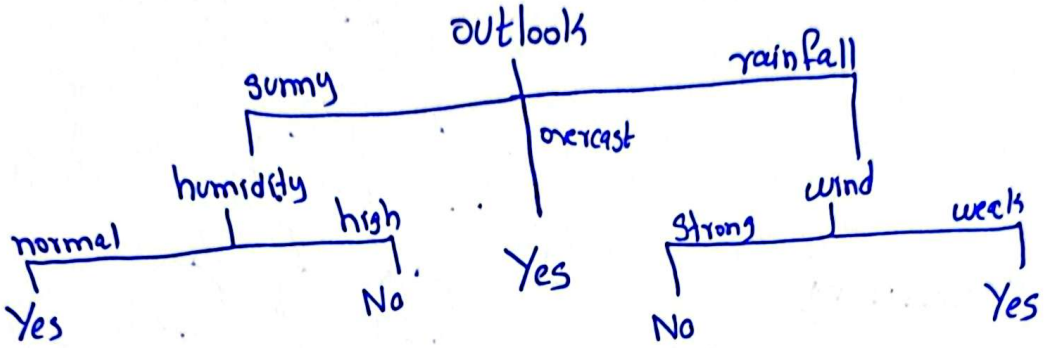
Gini (outlook = rainfall and wind = weak) = $1-(3/3)^2-(0/3)^2 = 0$

Gini (outlook = rainfall and wind = strong) = $1-(0/2)^2-(2/2)^2 = 0$

Gini (outlook = rainfall and wind) = 0



Now, Lets focus on sub data for strong and weak for wind and rainfall outlook feature



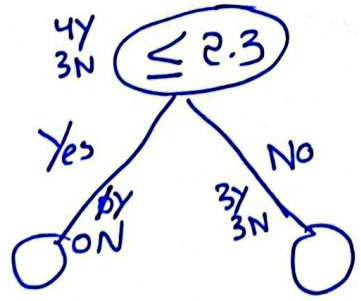
⟹ > Decision Tree Split on Numerical Features

f1	output
2.3	Yes
3.6	Yes
4	No
5.2	No
6.7	Yes
8.9	No
10.5	Yes

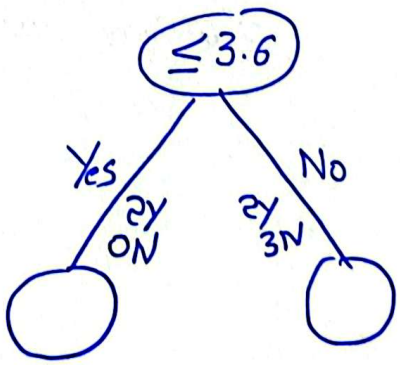
① Sort the value

② Select threshold

Threshold = 2.3

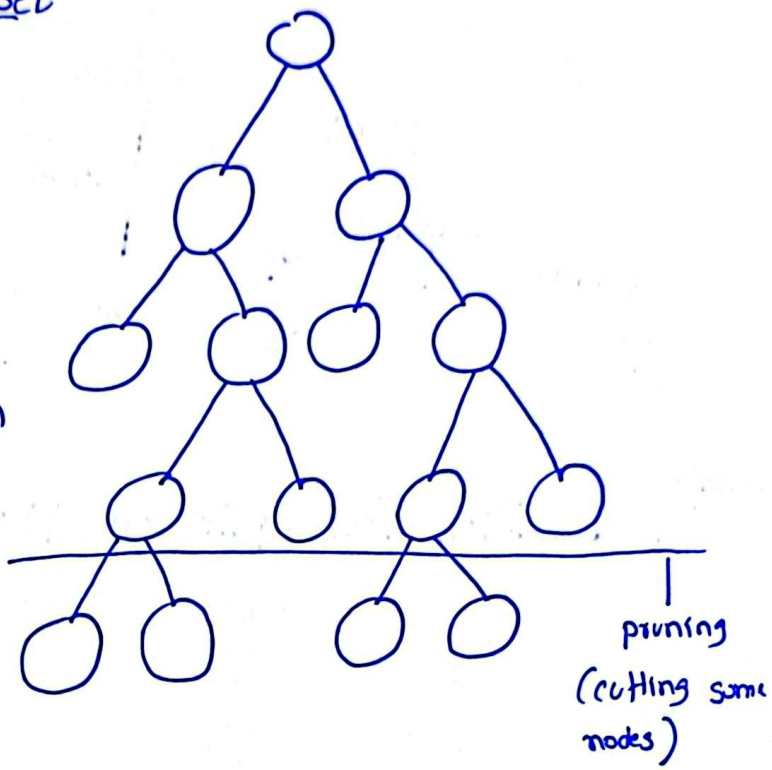


Threshold = 3.6



> Post Pruning and Pre Pruning

Training dataset



overfitting

1. Training Accuracy High
2. Test Accuracy Low
3. Low Bias
4. High Variance

Pruning → To reduce overfitting

Post Pruning

1- Construct decision tree
2- Prune it with respect to depth
3- Smaller dataset

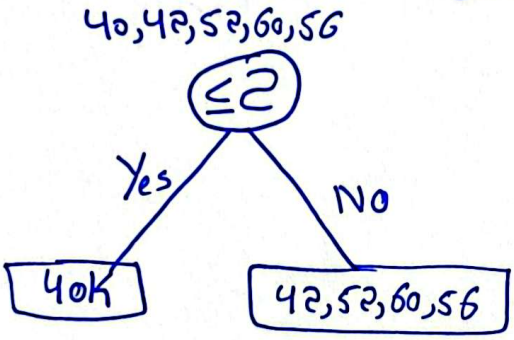
Pre Pruning

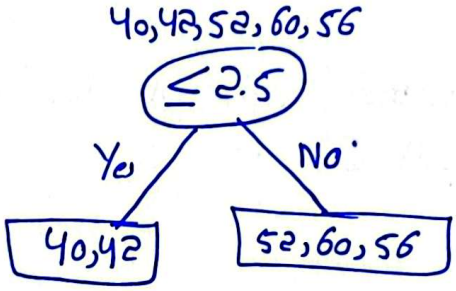
1- Play / Tune with hyperparameters

→ Decision Tree Regressor

Dataset

Exp	CareerGap	Salary
2	Yes	40K
2.5	Yes	42K
3	No	52K
4	No	60K
4.5	Yes	56K
		50K





Variance Reduction

$$\text{Variance} = \frac{1}{n}\sum_{i=1}^{n}(y-\bar{y})^2$$

$$\text{Variance of Root} = \frac{1}{5}\left[(40-50)^2+(42-50)^2+(52-50)^2+(60-50)^2+(56-50)^2\right]$$

$$= \frac{1}{5}\left[100+64+4+100+36\right]$$

$$= 60.8$$

$$\text{Variance of C1} = \frac{1}{1}\left[40-50\right]^2$$

$$= 100$$

Variance of C2 $= \frac{1}{4}\left[(42-50)^2+(52-50)^2+(60-50)^2+(56-50)^2\right]$

$= \frac{1}{4}\left[64+100+36+100\right]$

$= 51$

for splitting ①

Variance (Root) = 60.8

Variance (C1) = 100

Variance (C2) = 51

Variance Reduction

$Var(Root) - \sum w_i Var(child)$

$= 60.8 - \left[\frac{1}{5}(100)+\frac{4}{5}(51)\right]$

$= 60.8 - [20+40.8]$

$= 0$

Variance reduction for split ① is 0

Variance of Root-2 = 60.8

Variance (C1) $= \frac{1}{2}\left[(40-50)^2+(42-50)^2\right] = \frac{1}{2}[100+64] = 82$

Variance (C2) $= \frac{1}{3}\left[(52-50)^2+(60-50)^2+(56-50)^2\right]$

$= \frac{1}{3}[4+100+36] = 46.66$

Variance Reduction for split ②

$= 60.8 - \left[\frac{3}{5}\frac{2}{5}(82)+\frac{3}{5}(46.66)\right]$

$= 60.8 - [32.8+27.996]$

$= +0.004$

Variance Reduction for split ① is less than split ② so split ② will be selected as root node.